Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 10 (2011) 511 – 516

Procedia Environmental Sciences

2011 3rd International Conference on Environmental Science and Information Application Technology (ESIAT 2011)

A Technique of Filtering Dirty Data Based on Temporal-Spatial Correlation in Wireless Sensor Network

Chun-Hua Zhou[a], Bing Chen, Yong Gao, Chao Zhang Zhan-Jie Guo,a*

[a]*Zhengzhou Institute of Information Science and technology, Zhengzhou, 450001, China*

Abstract

The problem of filtering dirty data in wireless sensor network is studied in this paper. To solve the problem, the technique of filtering dirty data based on temporal-spatial correlation is proposed. The technique takes advantage of the temporal-spatial correlations of sensed data, and builds a temporal-spatial correlation model. The error data, namely dirty data which are produced by sensor nodes, are filtered out through the model and are divided into temporary bad data and permanent bad data. First of all, local nodes carry out the first-time filtering to filter the temporary bad data through temporal correlation. Secondly, second-time filtering is performed to filter the permanent bad data in a cluster through spatial correlation. Sensor nodes of every cluster are alternate to be the head nodes to balance the energy consumption, thereby reducing the death rate of the nodes to extend the life cycle of the whole network.

Keywords: Wireless sensor network; Dirty data; Temporal-spatial correlation

1. Introduction

Wireless Sensor Networks, short for WSN, are integration of sensor techniques, nested computation techniques, distributed computation techniques and wireless communication techniques. They can be used for testing, sensing, collecting and processing information of monitored objects and transferring the processed information to users [1, 2]. Wireless sensor network is a data-centric network [1]. Users can get interested data by data query. However, because of the shortcomings such as low accuracy, limited hardware resources, fragile anti-disturbance, sensor nodes would produce errors as a result of environment noise, hardware disturbance, environment temperature, lost of energy and hardware failure.

* Corresponding author. Tel.: +1-567-061-4886.
E-mail address:zchgjb@126.com.

doi:10.1016/j.proenv.2011.09.083

Data with errors will have a disgusting influence on the accuracy of the query results and waste lots of human and material resources.

Bad data generated in WSN is called dirty data [3~5]. On the one hand the accuracy of sensing data will be reduced because of the noise of the surrounding environment, interference caused by self-hardware, environment temperature, on the other hand failure sensor nodes caused by energy exhausted or hardware damaged will produce abnormal error data. According to the characteristics and nature of bad data, bad data is divided into two types: temporary bad data and permanent bad data [3].

In Wireless Sensor Networks, the data generated by sensor nodes has some certain temporal and spatial correlation characteristics [3, 6]. Temporal correlation means that there is a quantitative function relation exists between the data in the present moment and the data in the next moment. eg: As the temperature within a day in forest changes slowly, then in a relatively short time interval, the change of the sequential sampling data will be very small, possibly even the data remains invariable. Spatial correlation is that a quantitative function relation exists between the data generated by the nodes within a certain space. eg: In the same office, the temperature values perceived by sensor nodes will satisfy a certain error range [7].

Aiming at the problem of dirty data filtering in wireless sensor networks, a technique of filtering dirty data based on temporal-spatial correlation is proposed in this paper with the temporal - spatial correlation of the sensing data. This technique filters the dirty data twice with the temporal - spatial correlation: First, local nodes carry out the first-time filtering to filter temporary bad data with temporal correlation. Secondly, second-time filtering is performed to filter permanent bad data in a cluster with spatial correlation. The division technology of full-connected cluster is adopted to divide clusters [8], Sensor nodes of every cluster are alternate to be the head nodes to balance the energy consumption, thereby reducing the death rate of the nodes to extend the life cycle of the whole network.

In this paper, the related knowledge of wireless sensor network is introduced firstly. Section 2 is dedicated to related work. Problem definition, the specific problem solved and network model are given in Section 3; Section 4 describes the processing technology in detail, includes relevant concepts and concrete technology. Experiment analysis is in Section 5. Section 6 concludes the paper.

2. The technique of filtering dirty data based on temporal-spatial correlation

2.1. Filtering process of dirty data

Data filtering technique based on temporal-spatial correlation mainly adopts twice filtering technology, the first-time filter filters temporary bad data through temporal correlation; the second –time filter filters the permanent bad data through the spatial correlation. The filtering process is as follows:

Step 1. Sensor nodes sense data with some frequency.

Step 2. Divide the sensing data from step 1 into normal data and abnormal data by temporal correlation. For the normal data, it is transmitted directly. For the abnormal data, it is divided into temporary bad data and permanent bad data. If it is temporary bad data, filter the data directly, otherwise go to Step 3.

Step 3. Filter the data from Step 2 secondly to filter out the permanent bad data that can be judged by the proportion of the number of nodes which generate the abnormal data in the cluster.

Step 4. Transmit the normal data to the base station.

Table 1 defines some commonly used symbols in this paper.

Table 1. Table of symbols

Symbol	Definition
NQ	Queue of normal data
AQ	Queue of abnormal data
W	Sliding window
L	Length of sliding window
center	Center value of sampling data in W
ϕ	Fluctuation range fluctuation range
counter	Counter of abnormal data
λ *(0<λ<1)*	Threshold of percentage of valid data in AC
n	Count of nodes in a cluster
invalid_count	Count of nodes which generate the permanent bad data in a cluster
k	threshold of proportion of nodes which sense abnormal data in a cluster
k'	proportion of nodes which sense abnormal data in a cluster

2.2. The first-time filtering

In the process of the first-time filter, according to the characteristics that sensing data meets temporal correlation, Average Compare (AC) is adopted in this paper to determine whether or not the nodes sense the dirty data. The first-time filtering is divided into two phases: initialization phase and filtering phase. In the initialization phase, Sensor nodes sense data with some frequency, and put them into window W. When W is full, we calculate the average AVG (L). The difference between the average and the data in window is defined as follows:

$$\text{distance}(\text{AVG}(L), v_i) = |\text{AVG}(L) - v_i| \quad (1)$$

Here, v_i is the data sensed at the moment i. The function values are sorted from small to big. Then we select the samples which the first $\lambda * L$ distance values corresponding to as the sample. *center* is initiated as follows:

$$center = \text{AVG}(\lambda * L) \quad (2)$$

So far, the initialization phase is completed, as shown in Table 2.

Table 2. The algorithm of initialization

Algorithm 1: initialize()

Input: v_i
Output: *center*

```
1.  while( window isn't full && dataReady(v[i]))
2.    insert v[i];
3.  center = AVG(window);
4.  for every element v_i in the window
5.    dis[i] = distance(center,v_i);
6.  sort(v,dis);
7.  the first λ*WL elements in the sorting list make newWindow ;
8.  center = AVG(newWindow);
```

Because of the previous assumptions that the data values consider to be approximately equal in continuous short period, therefore, according to our daily experience and the basic cognition of the

environment for wireless sensor network, we can think that in normal circumstances new sampling values should fall into the region [center-φ, center + φ], the data which fall outside the region are regarded as dirty data.

In the filtering phase, sensor nodes judge whether W is full after collecting data, if W is full already, then calculate AVG (L) as the new *center*, save the data in W, and then empty W. Then, if distance(AVG (L), v_i)< φ then counter=0, and add v_i into NQ, otherwise counter = counter - 1, add v_i into NQ and clean AQ, that is filtering out the temporary bad data; else counter = counter + 1, and add v_i into AQ. Because the frequency of the temporary errors occurring is very low, so counter will be 0 after sampling just a few times.

The first-time filtering algorithm is shown in Table 3.

Table 3. The algorithm of first-time filtering

Algorithm 2: first_ time_filter()	
Input: v_i Output: NQ and AQ	
1.	initialize();
2.	while(dataReady(*v*))
3.	if(window is full){
4.	*center* = *AVG*(window);
5.	save window to NQ and clean window;}
6.	if (distance(*v*,*center*)<=*φ*)
7.	if(*counter* ==*0*)
8.	Insert(*v*,*W*);
9.	else{
10.	*counter*--;
11.	Insert(*v*,*W*);
12.	Clean AQ;}
13.	else{
14.	*counter*++;
15.	Insert(*v*,*AQ*);
16.	}

Table 4. The algorithm of second-time filtering

Algorithm 3: second_time filtering()	
Input: AQ Output: NQ	
1.	invalid_count = 0; i = 0;
2.	while (i <= n)
3.	if receive an AQ
4.	invalid_count ++;
5.	k' = invalid_count/n;
6.	if (k'<k){
7.	clean all AQ and notify the corresponding nodes;
8.	Send NQ;
9.	}
10.	Else
11.	Send abnormal event immediately

\When the permanent error occurs, such as sensing devices of nodes failure or fire happened suddenly, the temperature has been 100°or more than 100°, in the next sampling, counter will continue to become bigger, which means that the exception happened. If this happens, the resulting data is not immediately sent to the base station. The data of AQ will be transmitted to cluster heads to be filtered secondly.

2.3. The second-time filtering

After the first-time filtering, we just only filter out the temporary bad data, not the permanent bad data. Because we can't distinguish the abnormal data is the permanent bad data or produced by events happened in the first-time filtering. Therefore, in order to further thoroughly filter out the dirty data to reduce the amount of data and maximize energy savings, we adopt the second-time filtering, which is based on spatial correlation.

Cluster head receives data packets from nodes within the cluster, including normal data queue NQ and abnormal data queue AQ, and then it determines whether the data in AQ is permanent bad data, if so,

filters out it, no longer uploads; otherwise, it thinks that events happen, and needs to be reported immediately.

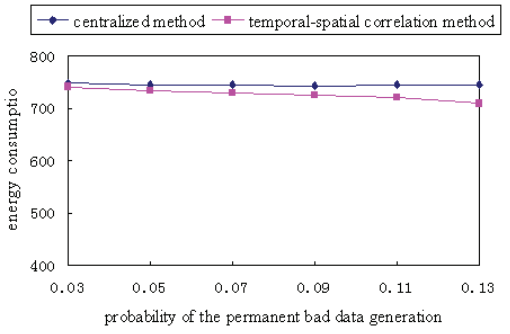
Given a threshold k, if k' < k, consider this abnormal data as dirty data; otherwise, think that events happen in this area. The process of second-time filtering is as follows: cluster head initializes that invalid_count = 0. When cluster head receives AQ, invalid_count = invalid_count +1. k' = invalid_count/n, if k' < k, consider this abnormal data as dirty data, and clean all received AQ, and notify the corresponding nodes to be failure; otherwise, think that the event happens in this area and report it.

The second-time filtering algorithm is shown in Table 4.

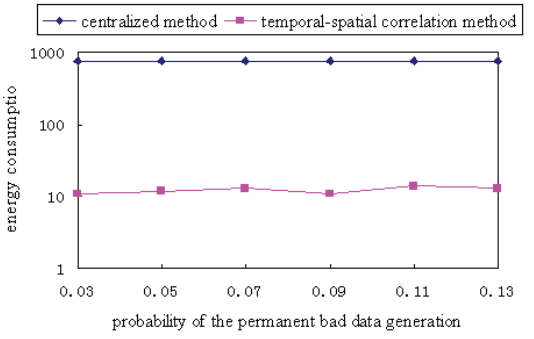
3. Experiment

In this section, we verified the effectiveness of the filtering technique proposed in this paper by experiments. We modeled 150 sensor nodes and produced a synthetic datasets to simulate the method.

Assume that the probability of the temporary bad data and the probability of the permanent bad data generation are 0.06, respectively. We respectively examine the influence of the temporary bad data and the permanent bad data on energy consumption as shown in Fig 1 and Fig 2. We compare energy consumption of centralized method and temporal-spatial correlation method in two conditions. The first condition: transmit all the correct data; the second condition: transmit the data when exception happened. Assume that the probability of occurrence of exception is 0.003.

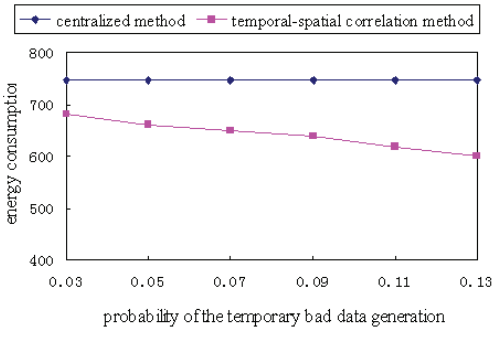


(a) The first condition

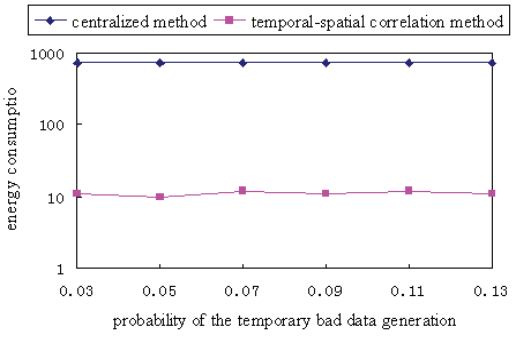


(b) The second condition

Fig 1. Effect of the permanent bad data on energy



(a) The first condition



(b) The second condition

Fig 2. Effect of the temporary bad data on energy

As shown in Fig 1 and Fig 2, the generation probability of the temporary bad data and the permanent bad data has almost no influence on centralized method in two conditions, but has certain influence on temporal-spatial correlation method. And in the second condition, using temporal-spatial correlation technique can greatly reduce energy consumption.

4. Conclusions

Dirty data filtering is the key work in the research field of WSN. We mainly study the dirty data filtering technique and propose a technique of filtering dirty data based on temporal-spatial correlation which carry out the first-time filtering based on temporal correlation and the second-time filtering based on spatial correlation to filter out the dirty data as much as possible. This technique can reduces the data transmission to save energy, prolong the lifecycle of network, and ensure the accuracy of the data transmission at the same time. We prove the feasibility and effectiveness of the technique by experiments.

References

[1] Tilak S, Abu-Ghazaleh N B, Heinzelman W. A taxonomy of wireless micro-sensor network models [J]. Mobile Computing and Communications Review, 2002, 1(2):1-8.

[2] Li J Z, Li J B, Shi S F. Concepts, Issues and Advance of Sensor Networks and Data Management of Sensor Networks [J]. Software Journal,2002,Vol.14:1717-1727.

[3] Z B, Gu Y, Li F F, Yu G et at. A Efficient Real-time Technique of Event Detection based on Temporal-Spatial Correlation Model in Wireless Sensor Network .Computer Research and Development(suppl) , 2006, 43(Suppl.):70~74.

[4] Ould-Ahmed-Vall E., Riley G., Heck B. A Geometric-Based Approach to Fault-Tolerance in Distributed Detection Using Wireless Sensor Networks [C]. In Proceedings of the IPSN, 2006.

[5] Chu D., Deshpande A., et al. Approximate Data Collection in Sensor Networks using Probabilistic Models [C]. In Proceedings of the ICDE, 2006, pages 48-60.

[6] Jeffery S. R., Alonso G., et al. A Pipelined Framework for Online Cleaning of Sensor Data Streams [R]. 2005, Report No. UCB/CSD-5-1413.

[7] Considine J, Li F, et al. Approximate aggregation techniques for sensor databases [C]. In Proceedings of the Int'l Conf. on Data Engineering: IEEE Computer Society, 2004, pages 449-460.

[8] Zhou C H, Yang X C, Wang B et al. A Processing Technique of Window-Based Approximate Continuous Queries in Wireless Sensor Networks [J]. Computer Research and Development(suppl) , 2006 Vol.43: 143-147.